2012

Deciphering a global network of functionally associated post-translational modifications

Pablo Minguez
European Molecular Biology Laboratory

Luca Parca
University of Tor Vergata

Francesca Diella
European Molecular Biology Laboratory

Daniel R. Mende
European Molecular Biology Laboratory

Runjun Kumar
Washington University School of Medicine in St. Louis

See next page for additional authors

Molecular Systems Biology 8; Article number 599; doi:10.1038/msb.2012.31
Citation: ***Molecular Systems Biology*** **8:599**

www.molecularsystemsbiology.com

molecular systems biology

Deciphering a global network of functionally associated post-translational modifications

Pablo Minguez[1], Luca Parca[2], Francesca Diella[1,3], Daniel R Mende[1], Runjun Kumar[4], Manuela Helmer-Citterich[2], Anne-Claude Gavin[1], Vera van Noort[1] and Peer Bork[1,5,*]

[1] Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany, [2] Department of Biology, University of Tor Vergata, Rome, Italy, [3] Molecular Health GmbH, Heidelberg, Germany, [4] Washington University, St Louis, MO, USA and [5] Max-Delbruck-Centre for Molecular Medicine, Berlin-Buch, Germany
* Corresponding author. Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. Tel.: +49 6221 387 8361; Fax: +49 6221 387 517; E-mail: bork@embl.de

Received 27.1.12; accepted 4.7.12

Various post-translational modifications (PTMs) fine-tune the functions of almost all eukaryotic proteins, and co-regulation of different types of PTMs has been shown within and between a number of proteins. Aiming at a more global view of the interplay between PTM types, we collected modifications for 13 frequent PTM types in 8 eukaryotes, compared their speed of evolution and developed a method for measuring PTM co-evolution within proteins based on the co-occurrence of sites across eukaryotes. As many sites are still to be discovered, this is a considerable underestimate, yet, assuming that most co-evolving PTMs are functionally associated, we found that PTM types are vastly interconnected, forming a global network that comprise in human alone >50 000 residues in about 6000 proteins. We predict substantial PTM type interplay in secreted and membrane-associated proteins and in the context of particular protein domains and short-linear motifs. The global network of co-evolving PTM types implies a complex and intertwined post-translational regulation landscape that is likely to regulate multiple functional states of many if not all eukaryotic proteins.
Molecular Systems Biology **8**: 599; published online 17 July 2012; doi:10.1038/msb.2012.31
Subject Categories: metabolic and regulatory networks; signal transduction
Keywords: post-translational modifications; protein regulation; proteomics; PTM code; PTM crosstalk

Introduction

After translation, protein function is mainly regulated by the tight interplay between protein–protein interactions and post-translational modifications (PTMs) (Seet *et al*, 2006). As many as 435 different PTM types are listed in the Uniprot database (The UniProt Consortium, 2010). They affect a significant fraction of eukaryotic proteins (Cohen, 2000; Weinert *et al*, 2011) and can interplay within or between these, either regulating different activities or mediating functional associations among modified residues of the same or different PTM types (Yang, 2005; van Noort *et al*, 2012). This PTM co-regulation has been shown to be crucial for the control of several important cellular processes (Kontaki and Talianidis, 2010). The examples of histone tail modifications (Latham and Dent, 2007) and p53 regulation (Brooks and Gu, 2003) have been used to suggest that most eukaryotic proteins may be subjected to PTM co-regulation to fine-tune their functional roles and a general 'PTM code' was postulated (Benayoun and Veitia, 2009). The conservation status of the respective sites, i.e., the modified amino acids, has been used as a proxy to measure PTM activity (Boekhorst *et al*, 2008), their location within protein structure (Landry *et al*, 2009), amino-acid specificity (Chen *et al*, 2010) or distribution in the quaternary structure (Jensen *et al*, 2006). Both conservation of modifications and PTM co-regulation are key features to understand protein regulation and their impact on the global protein interaction network of a cell.

Several studies reported on the conservation of individual PTM types (Boekhorst *et al*, 2008; Tan *et al*, 2009; Kumar and Balaji, 2011) and recently two PTM types have been comparatively analyzed (Weinert *et al*, 2011). However, to understand global PTM co-regulation, their evolution and the mechanistic insights that help deciphering the PTM code, a consistent framework has to be developed that allows the integration of many PTMs in a systematic manner.

Indeed, most mechanistic studies of protein modifications are still based on single PTMs (Choudhary *et al*, 2009; Oppermann *et al*, 2009; Zielinska *et al*, 2010) and are performed on individual or small groups of proteins (Brooks and Gu, 2003; Latham and Dent, 2007). Only recently, large-scale analyses revealed crosstalk of two types of PTMs which compete for the same residue (Danielsen *et al*, 2011), but many other ways of PTM functional associations are conceivable (Hunter, 2007). Also, *in-silico* perturbations of acetylation sites (Lu *et al*, 2011) has been used to measure structural changes in other PTM types.

We collected a large data set of experimentally verified sites for 13 abundant PTM types in 8 eukaryotes to (i) compare the speed of evolution of the different PTM types by means of a novel algorithm that takes into account the phylogenetic relationships among the species with the conserved amino acid, and (ii) to investigate, using information theory, the co-evolution of the modified residues as a proxy for their functional association. Focusing on PTM co-evolution within proteins, we identified a total of 35 pairs of different PTM types that are likely to interplay (>50% of all pairs analyzed), consisting of >74 000 modified residues, suggesting a global network of functionally associated PTM types. Our predictions cover well-known types of PTM crosstalk, but also hint at many new functional associations, in particular in secreted and membrane-associated proteins. Moreover, we derive functional context for many cases of PTM type interplay by identifying domains in which the functional association preferably occurs and by detecting linear sequence motifs that appear to be linked to specific pairs of PTM types.

Results

A non-redundant compendium of PTM sites in eukaryotes

We compiled >420 000 experimentally verified PTMs of 89 different types from public data sets from a total of 2485 species. After a preprocessing step to remove sequence redundancies, a total of 115 149 distinct modified residues of 13 types from 8 eukaryotic species from human to yeast were used for further analysis (Figure 1; for details see Materials and methods and Supplementary Dataset 1). The by far most abundant modification type in the data set is phosphorylation with almost 93 000 sites followed by N-linked glycosylation with 8827 sites (Figure 1A). C-linked glycosylation with only 45 collected sites was used as cutoff as with fewer sites the statistics do not allow to significantly reveal co-evolution (see Materials and methods and Supplementary Figure 1). The respective residues were found in a total of 21 046 proteins, 17 562 of them are phosphorylated implying a mean of 5.3 reported phosphorylation sites per protein, followed by acetylation of 3432 proteins (2.2 sites per protein), while C-linked glycosylation is the least frequent modification type in the data set covering only 12 proteins (3.8 sites per protein). Most PTM types show a preference for a particular cellular location, functionality and protein region (Supplementary Figure 2).

Proteins in our data set are not only modified by several PTMs of the same type but frequently also by multiple types. Each protein contains on average a total of 4.5 modified residues from the 13 PTM types studied and 20.8% of the proteins in our data set contain two or more modification types (Figure 1B). While proteins have in general less PTMs than expected (if the total PTMs in the data set are randomly assigned to the same number of proteins), a few of them have far more PTMs than the random expectation, which suggest that many sites are still to be discovered (Supplementary Figure 3). Thus, as it is also unlikely that we retrieved all PTM sites that have been experimentally verified, the PTM density used here has to be considered as a vast underestimate (Cohen, 2000). The highly abundant phosphorylated and glycosylated (O- or N-linked) residues are found in combination with all other PTM types (Figure 1C). The majority of observed PTM co-occurrences within proteins do not compete for the same type of amino acid (Wang *et al*, 2010; Danielsen *et al*, 2011) implying mostly other mechanisms of functional association. Comparing the distribution of the number of PTM types over proteins to the random expectation (PTMs in data set randomly assigned to the same number of proteins), we found that proteins have less PTM types than expected, which again suggest the incompleteness of the data set (Supplementary Figure 3).

Differing conservation of PTM types within eukaryotes

We comparatively studied the conservation status of the 13 PTM types as the first step for understanding their functional relations and their co-occurrence within proteins. As experimental data are not yet covering all organisms comprehensively, we assume, as implemented in other algorithms for similar purposes (Chica *et al*, 2008; Malik *et al*, 2008; Biswas *et al*, 2010), that the conservation of the site can be a good approximation for the conservation of the PTM. Indeed, this approach has been used to distinguish between functional and non-functional phosphorylation sites (Gnad *et al*, 2007; Holt *et al*, 2009; Tan and Bader 2012) and a less-strict criterion, the overall conservation of the proteins, was applied to determine the age of the PTMs functionality (Choudhary *et al*, 2009; Zielinska *et al*, 2010).

Thus, we aligned orthologs from 55 completely sequenced eukaryotic genomes taken from the eggNOG resource (Muller *et al*, 2010b) to the proteins with experimentally validated sites (for details see Supplementary Figure 4 and Materials and methods) and used multiple alignments to develop a Residue Conservation Score (RCS). RCS is composed of two elements (Figure 2A): (i) the evolutionary spread of species that contain a conserved residues, captured by the maximum branch length (MBL) of any two species containing the same residue as the PTM site and (ii) the conservation within the taxonomic range determined by the common ancestor of all species containing the conserved sites, expressed as Residue Conservation Ratio (RCR), see Materials and methods. By multiplying MBL and RCR, we take into account highly conserved PTMs that are only found in a narrow taxonomic range (e.g., primates) and taxonomically widespread PTMs (like phosphorylation) that are less conserved. For algorithm performance tests see Supplementary Figures 5 and 6. As different proteins have varying rates of evolution (Huerta-Cepas *et al*, 2007), we normalize every PTM RCS value using RCSs of non-modified amino acids identical to the one in the PTM site across all orthologs in the alignment, whereby we distinguish between ordered and disordered regions and calculates the relative RCS (rRCS; see Materials and methods, Supplementary Figure 4 and Supplementary Dataset 2 for details).

As expected, ordered protein regions are generally significantly more conserved than disordered regions in our data set (Supplementary Table 1). However, a separate comparison of the global distribution of RCSs for modified versus

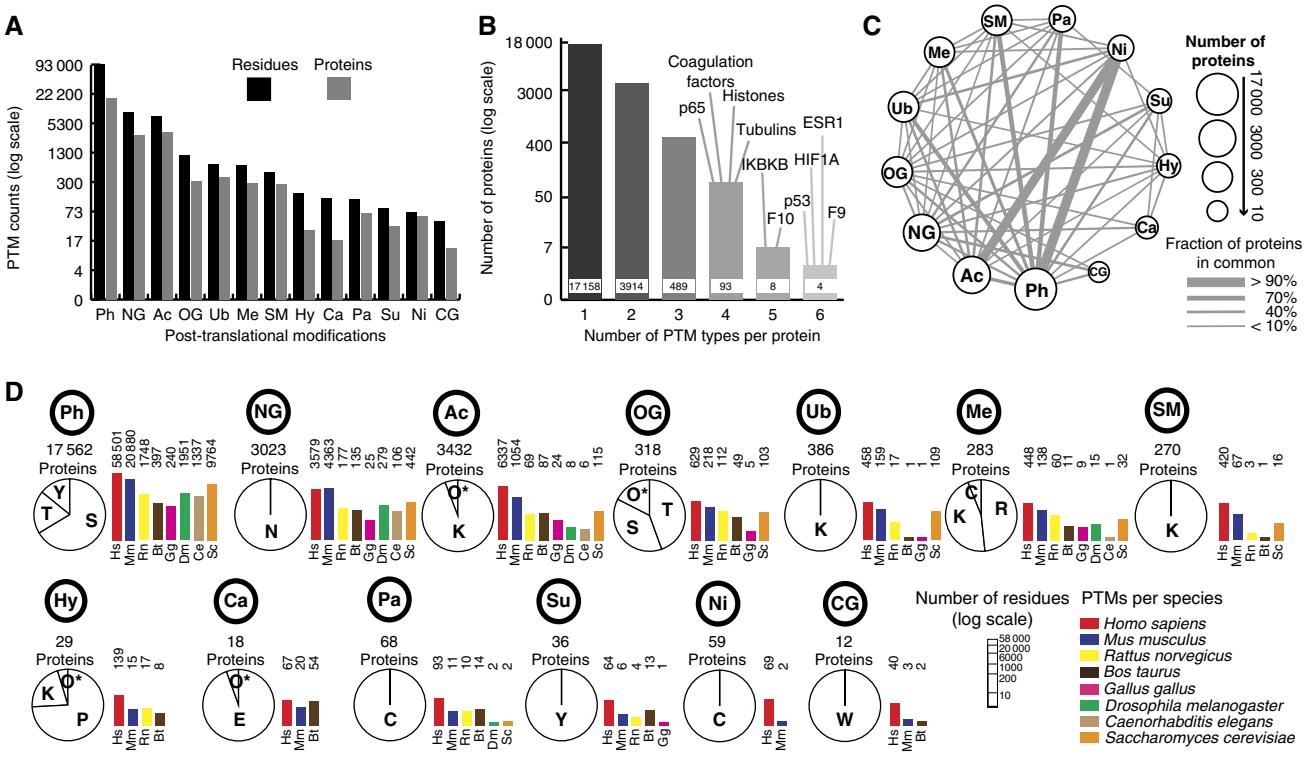


Figure 1 Statistics of the PTMs in the data set. (**A**) Number of residues and proteins per PTM type, PTM types are abbreviated as Ph (phosphorylation), NG (N-linked glycosylation), Ac (acetylation), OG (O-linked glycosylation), Ub (ubiquitination), Me (methylation), SM (SUMOylation), Hy (hydroxylation), Ca (carboxylation), Pa (palmitoylation), Su (sulfation), Ni (nitrosylation) and CG (C-linked glycosylation). (**B**) Breakdown of modified proteins by the number of PTM types per protein; the proteins with the highest PTM type frequency have all been intensively studied, e.g., coagulation factors, hypoxia-inducible factor or p53. (**C**) Co-occurrence of different types of PTMs within proteins, nodes size represent the abundance of proteins with a particular PTM type, the edge widths represent the number of proteins modified by the two respective PTM types normalized by the total number of proteins with the less abundant PTM type. Phosphorylated and glycosylated (O- or N-linked) residues are found in combination with all other PTM types followed by acetylation which it is not present together with C-linked glycosylation and carboxylation; only carboxylation and C-linked glycosylation (with the fewest sites in our data set) co-occur together with less than six other PTMs. (**D**) Breakdown of experimentally validated PTMs per species for each PTM type, the total number of proteins per PTM type, and the fractions of residues targeted by each PTM type (O* means others amino acids).

non-modified residues in both ordered and disordered regions reveals significantly more site conservation of all PTM types compared with background residues (Supplementary Figure 7 and Supplementary Table 2).

For example, if we choose a rRCS of 95 as a stringent cutoff to evaluate the conservation of a residue, meaning that the modified residue is more conserved than 95% of the non-modified residues, we find that 17.7% of the experimentally determined human phosphosites of serine are significantly conserved (16.5% for serine, threonine and tyrosine together). Using the same threshold, acetylations and N-linked glycosylations have higher rates of conservation, 26 and 20%, respectively. When comparing the 8 eukaryotes studied, human and mouse had very similar conservation levels while yeast sites appeared less conserved (Supplementary Table 3).

When comparing the different PTM types to each other using our scoring scheme (Figure 2B), carboxylation clearly stands out as the most conserved while SUMOylation is the fastest evolving PTM. Analysis of independent data sets derived from different species reveals that larger data sets (like human and mouse) show the best concordance while smaller ones contain more variability (Figure 2C), yet supporting the notion of similar roles of PTMs in different eukaryotic species (for more details see Supplementary Figures 8 and 9).

Co-evolution of sites within proteins reveals a global network of functionally associated PTM types

With the conservation status of the PTMs in hand, we were able to analyze the co-evolution of pairs of modified residues as a proxy of their functional association. The predicted functional associations (or the synonym 'interplay') are broadly defined here, ranging from physical interactions or competition for a site to co-regulation or involvement in at PTM signaling cascade. We use the loosely defined, but frequently used term crosstalk only in reference to published work.

We focus here on the co-occurrences of any two sites within a protein as the statistical framework is straightforward in contrast to co-occurrences in different proteins. We used mutual information (MI), corrected to exclude anti-correlation of residues, to measure the co-occurrence of two sites to identify pairwise co-evolution (Figure 3A). We evaluate the

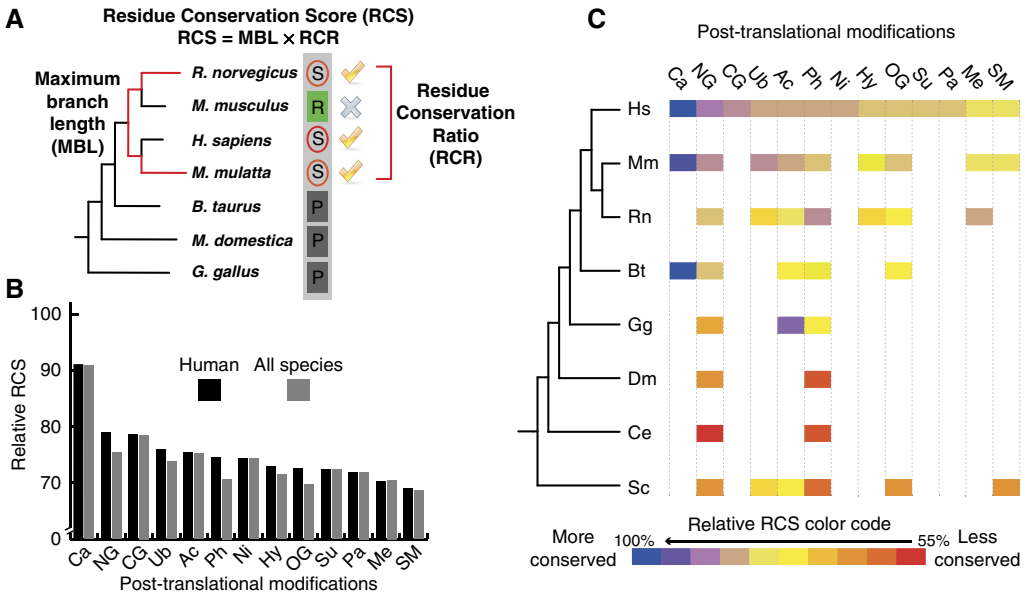


Figure 2 Differential conservation of PTM types. (**A**) The RCS is composed of two components: the MBL that is the longest evolutionary distance among species that contain a conserved modified residue and the RCR that quantifies the conservation ratio of the modified residue across the species in a taxonomic group in which a least one conserved site residue has been observed. The score is illustrated by a modified serine (circled) within a column of a MSA of orthologs where the species with the longest branch length containing the residue are *Macaca mulatta* and *Rattus norvergicus*; In the respective taxonomic group, 3 out of 4 species maintain the serine in the same position and thus the RCR is 0.75. (**B**) Average of the relative RCS (rRCS, obtained via comparison of RCSs of other residues in the protein, see Materials and methods for details) per each PTM type in human and in all species together. (**C**) Distribution of the mean of the rRCS in each PTM type across 8 eukaryotes. Colors indicate the degree of conservation (blue for more conserved residues and red for fastest evolving ones). PTM types are sorted according to the human values.

significance of the co-occurrence of two sites in the phylogeny of species encoding the orthologous groups (OGs) by permuting the species labels 100 times and then comparing the MI of the modified sites with the MI they would obtain in all possible tree scenarios (see Materials and methods and Supplementary Figure 10). MI has already been successfully used for analogous problems in molecular biology, such as the identification of interdependent mutations, protein interactions or protein residues co-evolution (Korber *et al*, 1993; Huynen, 2000; Martin *et al*, 2005) among others, and pinpoints here concrete pairs of PTM residues that are likely to be functionally associated within proteins.

A systematic analysis of the co-evolution of each pair of PTM sites within each protein of our data set (excluding those modifications that target the same residue) revealed that 74 386 residues in 10 325 proteins are co-evolving (Supplementary Dataset 3), i.e., predicted to be functionally associated, and this is only counting the pairs of PTMs for which the sites have been experimentally verified; the numbers increase >20-fold if one considers all orthologs in which the sites are conserved (1 683 187 sites in total). While 251 211 functional associations are predicted between residues of the same PTM type, 47 993 reveal co-evolution between different PTM types. All together, we find significant global co-evolution in 35 pairs of different PTM types while 12 PTM types (all except C-linked glycosylation) show also global co-evolution with themselves (links not shown). This implies a global network of predicted functional links among PTM types (Figure 3B) that involves in human alone 51 844 sites of 6013 proteins whereby many PTMs still have to be discovered implying considerably more intertwining of PTM types. This extensive interplay is in particular striking for some PTM types for which only few sites are known, suggesting that they must contain a high degree of co-evolution to be detectable with statistical significance (see details for robustness of the network in Supplementary Figure 11A).

All types of modifications but C-linked glycosylation, for which we have the fewest experimentally verified sites, are predicted to functionally associate on average with six other PTM types whereby competition of different PTM types for the same residues is not even considered here (see Materials and methods) and we only record co-evolution within the same protein. Phosphorylation, for which by far the most sites are known, globally associates with 11 other PTM types, closely followed by both major types of glycosylations as well as acetylation, which all are also PTMs with many known sites. Phosphorylation, N-linked glycosylation and carboxylation are found to have the highest fraction of co-evolution within a PTM type and differ from the other modifications in that they have more predicted functional associations within the PTM type than between PTM types (Supplementary Figure 11B and C). Although the total amount of co-evolution correlates with the abundance of known sites, some PTM types seem to be more often functionally associated than others: 82.4% of *O*-glycosylated proteins have at least one link while the corresponding fraction for phosphorylation is only 56.4% (Figure 3B). Of all the proteins with a recorded PTM, 37% contain at least one site that functionally associate with another PTM type, and another 61% contain associations between PTMs of the same type.

Pairs of detected co-evolving residues were significantly more conserved than random pairs of modified residues

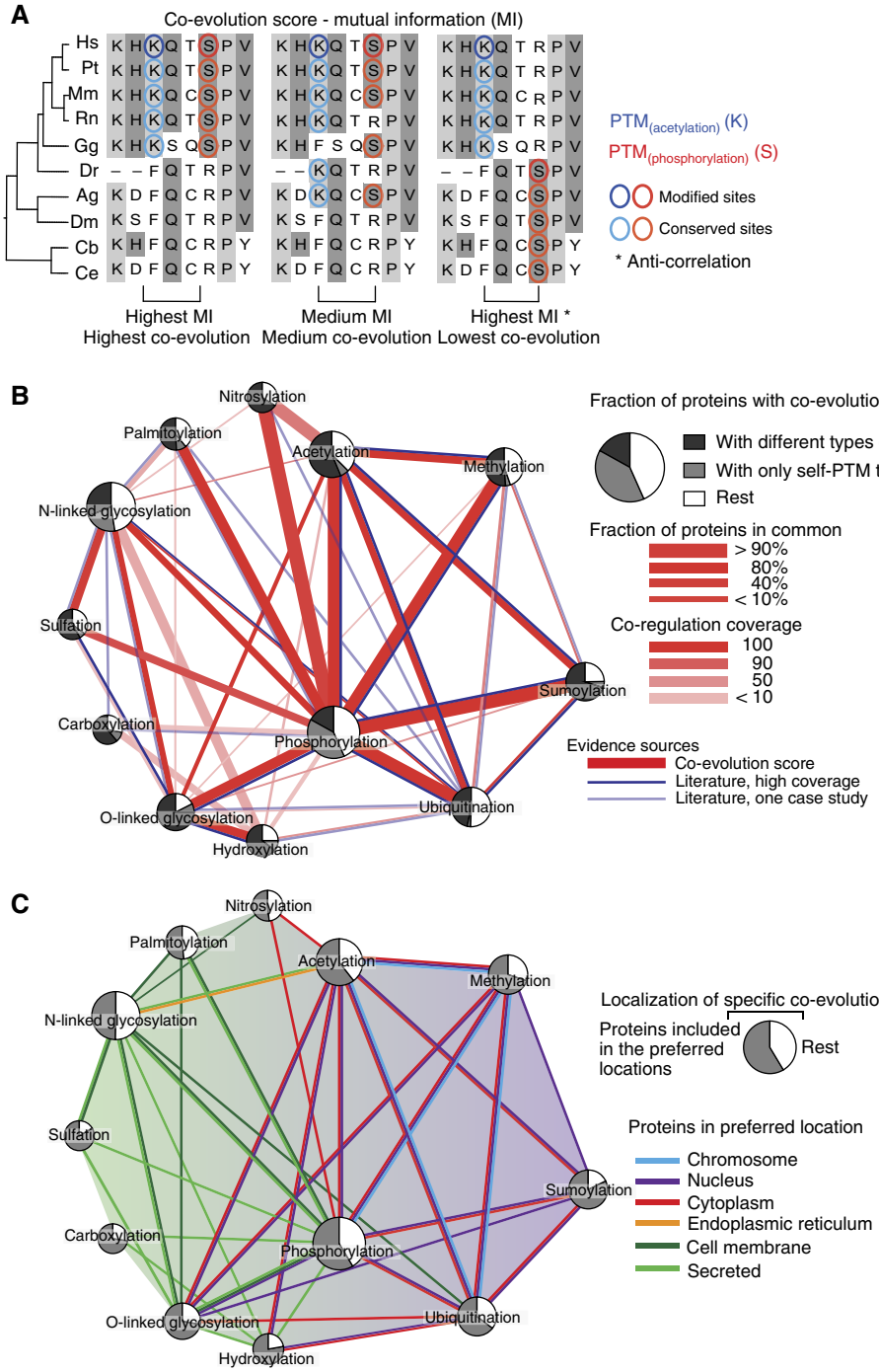


Figure 3 Global map of co-evolving PTMs. (**A**) Illustration of the co-evolving score based on MI: Two PTMs (acetylation in blue and phosphorylation in red) are pairwise evaluated in three different situations in which both residues are present in half of the orthologs from 10 species. The score includes the capacity of MI to address the level of dependency of two variables (maximum in both the right and left sequence alignments), and the degree of conservation of the amino acid across the species in the alignment (which is maximal in the left coupling). (**B**) Global network of co-evolving PTM types. PTMs types are represented as nodes whereby the size of the nodes indicate the number of proteins with such modification. Inside the nodes, the proportion of proteins with co-evolution is given for each PTM type: with the same (light gray) and with any other PTM type (dark gray). Red edges represent a predicted global functional association between two PTM types, the intensity of the color (from dark to light red) represents the fraction of the significantly co-evolving pairs of residues of all possible pairs within the proteins. The edge width indicates the fraction of proteins with the two co-evolving PTM types. Thus, thin dark red edges denote PTM pairs that have a low level of co-occurrence within proteins but if this happen they show a high level of co-evolution, thick light red edges link PTM pairs with a high co-occurrence in proteins of which only a few are co-evolving, yet significantly. The degree of association between PTM types (color intensity) does not always correlate with their frequency of co-occurrence in proteins (width). For example, although ubiquitination and SUMOylation are not frequently found to co-occur in proteins, almost in all of these instances they are predicted to be functionally associated. Furthermore, most nitrosylated proteins are also phosphorylated or acetylated but there is a much higher degree of co-evolution with phosphorylation. Blue edges represent known crosstalk between PTM types extracted from the literature, either well established (dark blue) or mentioned only in one case studies (light blue). (**C**) Network of co-evolving PTM types showing the preferred cell locations of the respective proteins.

(*P*-value $<2.2\ e^{-16}$; see Materials and methods) highlighting an underlying functional role and supporting the predictive potential of our approach. We next explored the functions to which each type of co-evolving PTM pair could contribute. When we compared the sets of proteins subjected to each of the co-evolving pairs of PTM types to sets of proteins subjected to the same type of PTMs which are not co-evolving, 43 out of the 47 pairs of PTM types found significantly co-evolving (35 between different PTM types and 12 self associations) were found to be enriched in particular functions or cellular components in the context of the Gene Ontology (Ashburner *et al*, 2000; Supplementary Datasets 4 and 5). In respect to known protein–protein interaction networks that they are part of, 23 out of the 47 pairs have a higher degree of connectivity than sets of random proteins and 10 out of these 23 if we compare against sets of proteins with the same PTM types which are not co-evolving (see Supplementary Figure 12 and Materials and methods for details).

To gain more functional insight into proteins with co-evolving PTMs, we extracted the preferred functionalities (more general functions than those defined by GO terms) that are annotated in the different pairs of PTM type interplay (see Materials and methods and Supplementary Figure 13). While some pairs of co-evolving PTM types target proteins that are involved in a broad spectrum of functions (eight of these pairs target proteins involved in more than four different general functions), other pairs of co-evolving PTM types are quite specific, i.e., 15 of them target proteins involved in a single function. Nuclear processes such as chromatin structure and dynamics, RNA processing and transcription are enriched in proteins with predicted functional associations between mainly phosphorylation, acetylation, methylation, SUMOylation and ubiquitination. However, each of the pairs have their own particularities, which again suggests the presence of a global PTM code that could be elucidated only with detailed mechanistic knowledge about each of these processes. We observe also that certain PTM types are linked to very different functionalities depending on which other PTM type they are co-evolving with. For instance, proteins that harbor co-evolving O-linked glycosylation and N-linked glycosylation sites appear to be involved in processes related to extracellular structures and signal transduction, while proteins harboring co-evolving O-linked glycosylation and ubiquitination, SUMOylation or methylation are linked very specifically to transcription. The latter can be still functionally separated at a more fine-grained resolution: e.g., proteins harboring co-evolving O-linked glycosylation and SUMOylation sites are specifically involved in cell differentiation and tissue development.

To assess the novelty within the delineated global network of functionally associated PTM types, we extracted Medline abstracts with co-occurring names of modifications (Supplementary Table 4) and reviewed the resulting large corpus manually (see Supplementary Materials for details on this revision). We then classified the pairs of PTM types as being (i) generally *known* (although we vastly extend the number of individual incidences), (ii) *proposed* (based on an individual protein or an individual case study so that almost all of our individual predictions can be considered novel) and (iii) *undescribed*. A total of 12 known PTM type crosstalks are well established in the literature and we see all of them in our network (Figure 3B). Furthermore, we generalize in 9 of 12 *proposed* PTM type crosstalks from individual observations or proposals to large number of instances. Finally, we identified 14 previously *undescribed* pairs of functionally associated PTM types for which to the best of our knowledge interplay has not been reported yet (solely red lines in Figure 3B).

As the network of functionally associated PTM types is strikingly enriched in membrane-associated or secreted proteins, we also systematically recorded the preferred cellular localization of the proteins in which we find co-evolution (Figure 3C). PTM type interplay appears to be enriched in proteins that can be classified according to the following preferred localizations: (i) In the nucleus or cytoplasm PTM interplay is found, e.g., in the histone regulation or well-studied crosstalk between phosphorylation, acetylation and ubiquitination occurs in a number of cytoplasmic processes such as protein degradation. (ii) In membrane-associated and secreted proteins where we found a considerable amount of unexpected functional associations. In human alone, 34 proteins that are at least partially secreted contain co-evolving and thus predicted functionally associated phosphorylations and N-linked glycosylations. Phosphorylation, associated with intracellular processes, was also found to be heavily co-evolving with palmitoylation and sulfation, two other modifications that occur exclusively in secreted proteins. As phosphorylation cannot actively happen outside the cells, it is likely that the interplay takes place in the endoplasmic reticulum in the context of decorating proteins with complex PTM structures before export. In total, 17 out of the 35 pairs of different PTM types with significant co-evolution levels occur mostly in secreted or membrane-associated proteins, implying that PTM interplay is crucial for the regulation of the protein export process. Finally, there is also PTM type interplay (iii) in proteins with a broad localization spectrum, which stand for example for functional interactions between phosphorylation, acetylation, ubiquitination and O-glycosylation.

Functional implications of co-evolving PTM types

With the large data set of co-evolving PTMs in hand, we tried to explore some of the mechanisms underlying the complex network of functionally associated PTMs shown in Figure 3B and C. We followed three different approaches: First, we analyzed the pairs of co-evolving PTMs for their proximity in sequence and structure to get an indication whether they could be physically interacting. It has previously been shown that some PTMs and PTM types can form clusters of sites that act as regulatory centers, e.g., the highly modified cassette of amino acids in p53 (Brooks and Gu, 2003). To generalize such physical interactions to all PTM types, we performed a non-parametric comparison of the distances between the modified residues in both sequence and structure (the former as a proxy of real spatial separation) with equivalent modified but not co-evolving residues from the same proteins. Our results indicate that 25 out of the 35 PTM type pairs identified as co-evolving have their residues closer in sequence or structure than comparable modified but not co-evolving residues (see significance levels at Figure 4A and B). Out of the PTM type

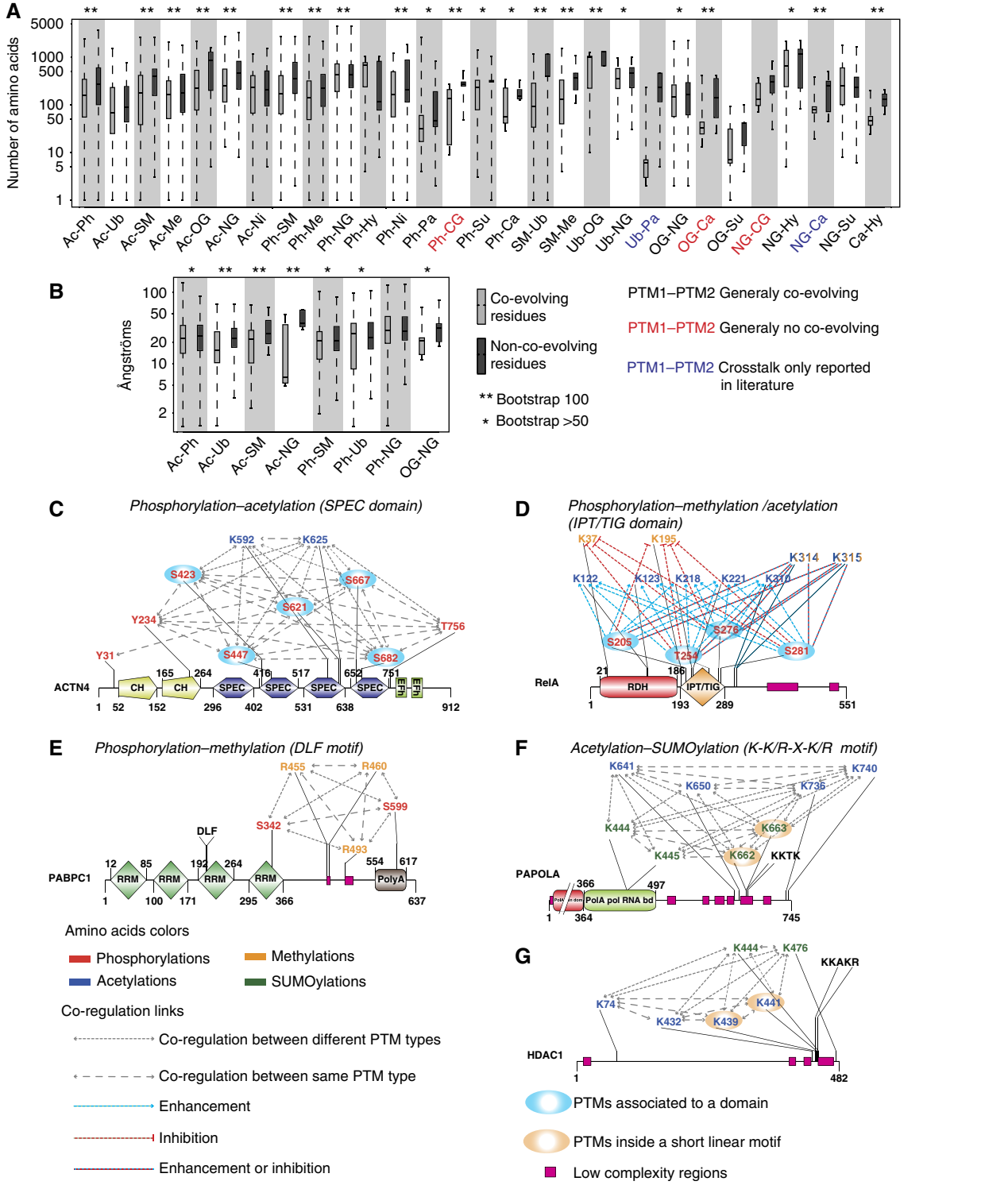
pairs that do not significantly co-evolve more than expected, we find five more pairs of PTM types with co-evolving residues closer in sequence than non-co-evolving modified sites. Two of these have some instances of crosstalk reported in the literature supporting the notion that our network of functionally associated PTMs within protein is far from complete. Nuclear and cytosolic PTM types such as those in proteasome degradation seem to be frequently interacting physically as they are found in close distance in sequence, in structure or in both. Also, residues with co-evolving PTM types associated to membrane-related or secreted proteins tend to be close in space, for instance carboxylated and hydroxylated residues known to co-regulate in coagulation factors and vitamin k-dependent proteins (Castellino *et al*, 2008), or N-linked glycosylated and acetylated sites which are found to be significantly closer in both structure and sequence in 41 proteins but their association is not reported in the literature (see Supplementary Material for a review on known and novel associations between PTM types).

In a second approach, we aimed to place PTM type co-evolution into a functional context via their preferred locations in protein domains. We performed an enrichment analysis to identify PTM types that are more likely located within a specific domain when they are co-evolving with a particular PTM type in comparison to their presence in isolation or their co-evolution with any other PTM type. A total of 31 Interpro domains were identified this way, 8 of them were already reported to be regulated by the functional association we predicted, 13 were described to be regulated by both PTM types independently, for 9 only one of the two PTM types had been associated with the domains and for one domain we did not find any literature connection to any of the two PTM types (Supplementary Table 5). An example in which the association of a domain with one PTM type has been previously shown but where we can add both mechanistic detail and generalization are spectrin repeats. Proteins containing co-evolved phosphorylation and acetylation sites are more often phosphorylated inside spectrin repeats than proteins that are either only phosphorylated or functionally associated to any other modification. Phosphorylation is known already to regulate several functions in which spectrin repeats play a role such as neuritogenesis (Bignone *et al*, 2007) or the stability of the erythrocytes membrane (Perrotta *et al*, 2001). Due to the enrichment and detection of the crosstalk in a number of spectrin repeats, e.g., in F-actin cross-linking proteins (Actn1, Actn4), Dystrophin (Dmd) and microtubule-actin cross-linking factor 1 (Macf1), we can hypothesize that the interplay between acetylation and phosphorylation in conjunction with the spectrin repeat might be a more general mechanistic scenario for the coordinated regulation of conditional binding of large cytoskeletal macromolecules and perhaps even for the formation of cytoskeletal networks. The prediction of mechanistic details is illustrated by Actn4 in which the acetylated residues have been identified by a global screen without mechanistic context (Choudhary *et al*, 2009): given the spectrin repeat enrichment, we hypothesize that it is their interplay with phosphorylation that fine-tunes the spectrin functionality (Figure 4C).

In a second example, we were able to infer a known regulatory process from the linkage of two different pairs of co-evolving PTM types to the same globular domain. It confirms our predictions and suggests particular residues to be involved that so far have not been related to the process. In the RelA subunit of the transcription factor NF-κ-B, we found an association of the cell surface receptor domain (IPT/TIG) with pairs of co-evolving phosphorylated and acetylated residues as well as with co-evolving phosphorylated and methylated residues. The activation of NF-κ-B via the regulation of its subunit RelA seems to be mediated by a series of crosstalk events between different modifications: e.g., acetylation has been found to inhibit methylation and to enhance ubiquitination that leads to protein degradation, thus methylation appears as a stabilizer that permits its DNA-binding activity (Yang *et al*, 2010). In addition to this regulation, phosphorylation of serine 276, which is inside the IPT/TIG domain, enhances the RelA acetylation (Chen *et al*, 2005) and phosphorylation also inhibits methylation (Levy *et al*, 2011). In our predictions, we found not only the reported S276 co-evolving with methylated and acetylated residues but also with three more phosphorylated sites inside the domain (S205, S281, T254). Those phosphosites would be co-regulating the protein localization together with a number of acetylated and methylated residues (Figure 4D). Our results suggest that the massive phosphorylation of the IPT/TIG domain would force RelA to degrade probably by means of enhancing a multi-acetylation (K122, K123, K218, K221, K310, K314, K315) and inhibiting methylation (Figure 4D). More examples of domain associated to co-evolving pairs of PTM types can be found in Supplementary Table 5. The domain approach illustrated here by the two examples thus provides mechanistic hypotheses and pinpoints to particular residues that are involved.

A third way to gain mechanistic insights is based on the occurrence of common short (3–10 residues long) linear motifs that are often found in unstructured regions. Proteins with a common interaction partner, or a family of interactors, can be enriched in such short-linear motifs that facilitate the binding (Sudol, 1998; Neduva *et al*, 2005). Using Gibbs sampling (Davey *et al*, 2010), we identified short-linear motifs for every group of proteins with a pair of co-evolving PTM types (see Materials and methods, Supplementary Table 6). We found 24 linear motifs, which can be merged into 12 clusters based on their sequence composition, that are enriched in proteins harboring 8 pairs of co-evolving PTM types; one of the motifs is shared by two different pairs of interplaying PTM types (so 13 relations in total). Based on a database similarity search (see Materials and methods) and a literature survey, we consider 4 out of the 12 motifs identified this way as being novel. In all, 7 of the 8 known motifs had been identified in the context of PTMs, so that we see our approach as confirmatory, although in each of these cases we also add novel aspects. As many as five motifs have been already associated to both PTM types individually but had never been connected to an interplay, two more motifs have been found to be associated to only one of the two co-evolving PTM types for which an enrichment was found and the last two known motifs have not been connected yet to PTMs (for details and references see Supplementary Table 6). For example, the known linear motif DLF (Asp–Leu–Phe) was found to be enriched in 19 proteins containing co-evolving methylated and phosphorylated residues. The DLF motif had previously been implicated in the

binding of CEBPB to the Rb–E2F complex, which is involved in cell-cycle regulation and synthesis of DNA (Darnell *et al*, 2003) although no PTMs were studied to regulate this process. We found a DLF motif in the methylated–phosphorylated protein P53bp1 which also interacts with the Rb–E2F complex (Mani *et al*, 2008) suggesting a similar, hitherto unknown regulation mechanism. DLF has also been associated to other cellular processes such as DNA replication, repair (Dalrymple *et al*, 2001) and vesicle trafficking (Mills *et al*, 2003) and we identified additional proteins from these processes harboring functionally associated methylated and phosphorylated residues. Furthermore, at least 6 of the 19 proteins in this data set, including PABPC1 (Figure 4E), are involved in RNA processing, identified by a functional enrichment analysis (see



Materials and methods), suggesting a novel functionality for the motif associated to co-regulated methylation and phosphorylation residues.

An example of how protein localization can be regulated by PTM interplay is the functional association between acetylated and SUMOylated residues within the nuclear localization signal (NLS), a short-linear motif (K-K/R-X-K/R) that triggers the signaling cascade leading to the nuclear translocation of proteins upon recognition by importins (Lange *et al*, 2007). We found this motif in 35 proteins harboring co-evolving acetylated and SUMOylated residues. A connection between phosphorylation and acetylation as such has already been described: While acetylation of NLS serves as a signal for nuclear import (Spilianakis *et al*, 2000), phosphorylation is a signal for cytoplasmic retention (Harrison *et al*, 2010). SUMOylation has also been reported as a signal to initiate the transport to the nucleus, e.g., for the Daxx protein (Chen *et al*, 2006); while other proteins first need the NLS signaling for their transport and are SUMOylated afterwards (Rodriguez *et al*, 2001). From the set of proteins with the (K-K/R-X-K/R) motif we found both mechanisms of regulation. On one hand, the protein Poly (A) polymerase α (PAPOLA) is SUMOylated at two of the lysines of the NLS motif (KKTK) that are co-evolving with several acetylated residues. SUMOylation of PAPOLA is required for its transport to the nucleus (Vethantham *et al*, 2008) while acetylation would change its role to inhibit the translocation (Shimazu *et al*, 2007), Figure 4F. In contrast, histone deacetylase 1 (HDAC1) has two acetylated lysines inside the NLS motif (KKAKR) that are co-evolving with two SUMOylated residues, suggesting that SUMOylation probably takes place in the nucleus after the protein transport (Figure 4G). More examples on associations of co-evolving PTM types to short-linear motifs can be found at Supplementary Table 6.

Discussion

Despite increasing efforts in large-scale discovery of individual PTM types, they are still an understudied source of cellular complexity (Deribe *et al*, 2010). Patterns of residues that are modified by different PTM types change during the life time of a protein, whereby little is known mechanistically about their interplay and dynamics in most cellular processes.

Here, we present the first large-scale study that comparatively analyses several PTM types in different eukaryotes, estimate their conservation in a consistent framework and derive a global network of functionally associated PTM types based on co-evolution of modified residues within proteins.

Given the incompleteness of data and the inherited underestimate of co-regulated sites (residues that do not co-occur might also be functionally associated given the fast evolution of some sites, so conservation is only one of the features that can be used to detect associations) we expect that the incidences of PTM interplay, each enriched in proteins with particular functionalities, will vastly expand in the future and should provide sufficient information to eventually decipher the proposed PTM code in a global way by adding more detailed information about the mechanistic nature of the functional associations.

The global network of functionally associated PTM types revealed already spatio-temporal specificities of PTM interplay (Figure 3C). Central to the PTM network appear phosphorylation, acetylation, ubiquitination and O-linked glycosylation that control both temporal events (e.g., transcriptional regulation) and processes that govern protein localization (e.g., export or membrane-association). These modifications seem widespread insight the cell and are amended by others with more restricted localization in secreted proteins or the nucleus. Although details of their predicted functional associations still need to be carved out, the tendency for close proximity in sequence and structure indicates a tendency for physical interactions of co-evolving PTMs in particular pairs of PTM types. Moreover, as we restricted our analysis to eukaryotes but prokaryotes harbor also a number of PTM types, it remains to be seen how far back in evolution this fundamental principle of protein regulation can be traced.

Most of the domains and motifs we found to be linked to functionally associated PTM types were already connected to individual PTMs or to a regulation of protein interactions; the role of PTM type interplay in these appears new as the regulation of protein binding has only been reported for individual PTMs (Neduva *et al*, 2005; Seet *et al*, 2006). Many more such links are likely due to the incompleteness of the data set. Our analysis provides first insights into the vast amount of interdependencies of different PTM types in the formation of distinct functional states of proteins from all cellular processes. It is likely that this global network also extends to PTM interplay between proteins, e.g., within protein complexes as already demonstrated at a small scale (Shukla *et al*, 2009). Furthermore, as we only analyzed pairwise interactions, examples of local, physically linked

Figure 4 Properties and functional implications of functionally associated PTMs. Residues of several co-evolving PTM types are found to be close in sequence (**A**) or structure (**B**) compared with equivalent modified, but not co-evolving residues in the same proteins. All PTM types pairs shown in (A, B) have co-evolving sites significantly closer than control sets with an adjusted *P*-value by FDR <0.05 in some of the 100 repetitions the analysis was repeated as we work with random sets as background. The bootstrapping values for the number of times the difference was found significant is showed by * (>50 times) and ** (100 times). The PTM types pairs are classified according to their status in our prediction, in black the pairs that we predict to be functionally associated, in red the ones that did not show a global co-evolution and in blue the pairs found to be associated only in the literature, in this two last cases even if a global co-evolution was not significant, several residues were found to be significantly co-occurring. For more distance analysis see Supplementary Figure 14. (**C**) Co-evolving PTM types are associated to protein domains probably regulating their activity, illustrated by phosphorylated and acetylated residues in the spectrin repeats of the protein ACTN4, spectrin repeats are in general enriched in the association between these two PTM types. (**D**) Protein RelA is phosphorylated inside the cell surface receptor domain in four residues (S205, T254, S276 and S281) that are found co-evolving with both, methylated and acetylated residues. As suggested by the reported co-regulation between phosphosite S256 and acetylated and methylated residues, a more general scenario is proposed where four phosphorylation would be enhancing the acetylation of seven lysines that would lead to protein degradation. In the absence of phosphorylation, four residues would become methylated and the protein translocated to the nucleus. (**E**) The DLF motif is associated with the co-evolution of phosphorylations and methylations, illustrated by the example of the PABPC1 protein. (**F**, **G**) Co-evolving methylated and SUMOylated residues can regulate protein localization in different ways depending on which type of modification is placed inside the nuclear localization signal (NLS) motif.

networks of PTMs (Figure 4C–G) illustrate that those local PTM networks within proteins will increase the number of functional states in a combinatorial way. Linking these to protein interaction networks would imply a complexity that would affect each cell types in different ways and allows fine-tuning of the regulation of all cellular processes.

Materials and methods

Data collection, annotation and sequence redundancy reduction

We compiled all PTMs available at UniProt (The UniProt Consortium, 2010), dbPTM (Lee *et al*, 2006), PHOSIDA (Gnad *et al*, 2010), PhosphoSite (Hornbeck *et al*, 2004), HPRD (Keshava Prasad *et al*, 2009), OglycBase (Gupta *et al*, 1999) and PhosphoELM (Dinkel *et al*, 2010). UniProt predictions labeled as inferred 'by similarity' or 'potential' were not included in the data set, nor in the analysis. For each modification, we recorded the protein id, type of modification, organism, amino acid and sequence position.

After collection, performed id and OG mapping and retrieved the protein sequences. We performed a pre-processing task to discard sequences that are identical or overlapping (e.g., peptides into protein sequences). The STRING database v8.3 (Jensen *et al*, 2009) is used for the id mapping and sequence retrieval. Sequences with no exact match to the corresponding sequence in the OG were excluded from the analysis. From the resulting number of PTMs and species present in the filtered data set, we selected only those sets, PTM types and organisms, with enough data to perform an accurate statistical analysis (see next point).

The selected proteins were annotated to their cellular locations according to Uniprot cellular component keywords ontology. The protein function was inferred from their OGs (Tatusov *et al*, 1997) as provided by the eggNOG database (Muller *et al*, 2010b).

Selection of PTM types to include in the analysis

In the collection of PTMs we found a large set of PTM types with only a few residues annotated in the databases. To assess the lower limit of the number of residues that a PTM type should have to be able to be detected significantly co-evolving with any other PTM type, we performed a simulation to test the capacity of our statistical framework to detect global co-evolution between two PTM types.

We first calculated the average ratio of modified residues versus non-modified residues (background) measured in all comparisons (0.13) and the average ratio of pairs of residues with significant co-evolution versus the pairs not found co-evolving for both sets: modified residues and non-modifies residues (1.53 and 0.45, respectively). Using these proportions we simulated the results (number of co-evolving residues versus not co-evolving residues in modified and non-modified residues) that are the input for the Fisher test that calculates whether a pair of PTM types is globally co-evolving compared with comparable non-modified residues. This type of analysis was performed for 50 simulated PTM types (each of them increases its number of residues in one, from 1 to 50 residues). We introduced two variables in these simulations: (i) the number of residues with other PTM type that are present in the same protein of at least one residue of the simulated PTM type (overlapping residues), we performed simulations for 1 to 8 overlapping residues and (ii) the number of not co-evolving modified residues of the simulated PTM type (from 0 to the number of residues overlapping).

Per each of the rounds of simulations (50 simulated PTM types with x number of overlapping residues and y number of not co-evolving residues), we generated the input for the Fisher test and added to the pipeline that gets adjusted P-values (by false discovery rate (FDR)) including all the results for the comparisons of the larger PTM types (13 in total) to simulate a real scenario in the adjustment of the P-values. See results in Supplementary Figure 1.

Evaluation of the number of PTMs and unique PTM types per protein

We built the expected distributions of the number of proteins with a particular number of PTMs and PTM types by randomly assigning the modifications we have in our data set to the proteins in each of the species in the study. We used the non-parametric Kolmogorov–Smirnov test to compare the expected distributions with the distributions extracted from our data set. We used Fisher exact test to evaluate whether the number of proteins with more than one PTM type in our data set is more than expected. See results in Supplementary Figure 3.

Functional enrichment analysis of protein with a specific PTM type

We analyzed the enrichment of protein functions and cellular location as well as type of protein region where the modified residues are located (i.e., ordered or disordered regions) for each set of proteins with a particular PTM type. The functional classifications for the proteins were obtained from metazoa OGs (meNOGs) from the eggNOG database, as a consensus between coverage and wealth of annotation. The reference set for the functional enrichment was the whole set meNOG functional classification. As mentioned before, we used Uniprot annotation for deriving protein location. We used the whole set of proteins from Uniprot from the 8 selected species from which we study the PTMs as the reference set for the location enrichment analysis. We also annotated the protein regions (ordered or disordered regions) where the modifications are placed and performed an enrichment analysis of any of these two categories in the global distribution of ordered and disordered region of the proteins belonging to the species under study. The DisEMBL algorithm v.1.4 (Linding *et al*, 2003) was used for the detection of disordered regions within proteins using COIL definition for the disordered regions.

For all the comparisons, we applied a Fisher exact test with P-values adjusted by FDR for the whole set of tests. Adjusted P-values below 0.05 were taken as significant.

Species tree

We built a phylogenetic species tree based on the NCBI taxonomy, which is known to be accurate for most taxa (Benson *et al*, 2010; Sayers *et al*, 2011), and inferred branch lengths. To assess the branch lengths, we generated alignments of 40 ubiquitous, single copy marker genes (Ciccarelli *et al*, 2006) for 853 different species using AQUA (Muller *et al*, 2010a) and combined the tree topology of the NCBI taxonomy tree with them using PhyML (Guindon *et al*, 2010). The resulting tree was manually curated and genomes that had an erroneous placement in the NCBI taxonomy tree were removed. The final tree includes 851 taxa including 35 eukaryotes, 43 archaea and 773 bacteria. Of the eukaryotic taxa included in the tree, 17 were metazoan (8 mammalia; 3 primates) and 13 were fungi.

Mapping proteins to OGs and tree generation pipeline

The database eggNOG v2.0 was used to map every protein in the data set to the oldest eukaryotic OG in which the protein is present and generated a multiple sequence alignment (MSA) and sequence tree using PhyML v.3.30. TreeKO (Marcet-Houben and Gabaldón, 2011) algorithm was used to root the tree and decompose the sequence tree into a set of all possible pruned trees with no duplications and in consonance with the species tree topology. The set of pruned trees includes all combinations of splitting the duplication events so that there are no trees with paralogous sequences but at least one tree for each paralog. The branch lengths for this tree are derived from the species tree. For all trees in this set that include the reference sequence, the residue RCS (see next section) is calculated for the modified residue. The tree in which the modified amino acid presents a higher score is selected for the evaluation.

RCS calculation

The RCS for amino acid aa (RCS_{aa}) evaluates the conservation of a particular amino acid (aa) within its position in the MSA of the corresponding OG. As explained in previous section, the resulting tree generated from the OG has been reconciled and has branch lengths information based on the species tree. The RCS_{aa} is composed by two components, the RCR, which represents the occurrence of the amino acid in the exact position in the sequences present in the pruned tree generated from the MSA and the MBL, which is the maximum branch distance between any pair of the species represented in the tree with the modified residue present in the aligned position.

To avoid that the size of the OG affects RCS, we get the oldest common ancestor of any two pairs of sequences with the amino acid conserved and only the descendants of this common ancestor are taking in consideration for the calculation. Thus, $RCS_{aa} = RCR \times MBL$ where $RCS_{aa} = \frac{N_{aa}}{Total_{Sequences}}$ and N_{aa} is the number of times aa appears in that position.

Reference distribution calculation for the normalization of the RCS

As the overall conservation status of the protein can indeed be a bias for the measurement of the conservation of the modified site, we generated a specific reference distribution of conservation of non-modified residues for every modification in the data set in order to normalize its RCS. To build the reference distribution for a particular modified site, we calculate the RCS for all non-modified residues from the OG that are of the same type of amino acid as the one with the PTM and it is placed in the same class of protein region (ordered or disordered). The RCS of the modified site is then mapped into the reference distribution to calculate the percentile of its value, this percentile is what we name relative RCS (rRCS). Only those modifications with >10 values in the reference distribution were selected for the conservation analysis.

Comparison of conservation distributions

We used the non-parametric Kolmogorov–Smirnov test to calculate the statistical significance of two distributions of RCSs or rRCSs.

In order to obtain a ranking of the overall conservation status of PTM types and function/location specific sets of PTMs, we calculate the mean of all rRCSs for every of these sets.

Extraction of co-evolving PTM pairs

We used the MI algorithm (Cover and Thomas, 1991) to evaluate the co-evolution of two PTMs in the same protein. The MI of two variables (Y, X) is measured as $MI(X;Y) = \Sigma_{x \in X} \Sigma_{y \in Y} p(x, y) \log \frac{p(x,y)}{P_1(x)P_2(y)}$ where $p(x, y)$ is the joint probability of X and Y, and $p_1(x)$ and $p_2(y)$ are the probabilities of X and Y, respectively. As MI measures the dependency of two variables, the PTM pairs modifying exactly the same residue within the protein were excluded from the analysis. We also excluded proteins which OG have <7 species and modified sites who are not conserved in at least 4 species and at most in the total number of species in the OG minus 3. MI values representing anti-correlation were converted to negative values to allow this measurement to distinguish between cases were co-evolution is real (common co-occurrences of sites in the same species) and cases were the presence of the site in the species set is complementary (high MI value but low co-evolution), see Figure 3A.

In every protein we built a set of different PTM pairs with all the modified residues present in the sequence. In order to avoid the effect of the possible co-evolution of PTMs of the same type increasing the global co-evolution measurement of a pair of PTMs due to multiple cross-evaluation, for a pair of PTM types, we only allow a particular modified residue to be present in a single pair of PTMs, the selection is done randomly. We then calculate the MI for the set of pairs selected and per each MI calculated we permute the species labels of one of the two residues 100 times and calculate again its MI. The permutations generate a reference distribution where the MI can be mapped in with a non-parametric approach. If the percentile of the modified residue is above 95 we assume the residues are co-evolving to a significant degree. The same approach is applied to a set of random pairs of non-modified residues selected under same conditions, same amino acids and same protein region distributions.

To evaluate the global co-evolution of two types of PTMs (as well as self co-evolution), we performed a Fisher exact test to see whether the ratio of the modified residues above 95 percentile on their reference distributions is significantly higher than the same ratio in non-modified residues. As the selection of the random pairs of residues can affect the result of the Fisher test, we repeated the whole process 100 times. *P*-values were corrected for FDR and value <0.05 taken as significant. The number of times we found that a pair of PTM types is significantly associated (maximum 100 and minimum 1) is then used to evaluate the coverage of the co-evolving pairs of PTMs within the whole set of pairs with the same type of PTMs.

Comparison of conservation between co-evolving and non-co-evolving modified residues

We compared by a Fisher exact test the number of times the rRCS of the modified and co-evolving residues is greater and lower than 95% against the same parameters for the modified but no co-evolving residues.

Preferred functionality for a set of proteins

To extract the preferred functions of the set of proteins with at least one pair of co-evolving modified residues, we used the annotation provided by the proteins OGs in the metazoa level (meNOGs) from the eggNOG database. A functionality is classified as preferred when the number of proteins annotated with such term is above the number of proteins annotated to any of the cellular locations present in the set as expected by chance plus 10 percent.

Preferred cellular location for a set of proteins

To extract the preferred location of the set of proteins with at least one pair of co-evolving modified residues, we used the cellular component annotation extracted from Uniprot (see above). A cellular location is classified as preferred when the number of proteins annotated with such term is above the number of proteins annotated to any of the cellular locations present in the set as expected by chance plus 10 percent.

Gene Ontology enrichment analysis

We performed a Gene Ontology terms enrichment analysis to every set of proteins with a particular pair of co-evolving PTM types using the FatiGO (Al-Shahrour *et al*, 2004) software available at the Babelomics suite (Medina *et al*, 2010). We first filtered the sets of proteins with co-evolving pairs of PTM types to include only human proteins (by far the most abundant in the data set). We did the same filtering to proteins sets with the same type of modifications but found not co-evolving and we used them as respective background lists. We evaluated the enrichment of Gene Ontology terms in biological processes, molecular function and cellular component categories restricted to a propagated annotation in level 5 (as a good compromise between the detail and the amount of the annotation retrieved) and using a two-tailed Fisher exact test for the enrichment analysis. *P*-values are adjusted by FDR and significant values are taken as <0.05.

Enrichment analysis in protein–protein interaction networks

We used the software SNOW (Minguez *et al*, 2009) available at Babelomics suite (Medina *et al*, 2010) to evaluate the topological parameters of the protein–protein interaction networks formed by the

set of proteins with a particular type of co-evolving pair of PTM types. As in the GO enrichment analysis previously described, we filtered the proteins sets to include only human proteins. We generated two types of background so the analysis was performed twice (all results shown), first as in GO enrichment analysis we generated the set of human proteins with same type of modifications but not co-evolving and the second type of background was offered by the SNOW software as a set of 500 random sets of proteins (with the same number of proteins as the query set). The minimal connection network (MCN) in both comparisons was generated allowing a single external protein to link proteins in the set and taking protein–protein interactions from the curated set SNOW provides. *P*-values are adjusted by FDR and significant values are taken as <0.05.

Analysis of the proximity of the co-evolving PTMs

We measured the distance between PTM sites as the number of intervening residues and in the available protein structures, in Ångströms, of every pair of co-evolving amino acids for each pair of PTM types. To have a reference distribution for each pair of PTMs, we randomly selected the same number of amino acids with the same PTM types, but found not co-evolving, from the same proteins and measure their distance in the same way. Both, the distribution of distances in the modified residues and the reference are then compared applying Kolmogorov–Smirnov test. *P*-values are adjusted by FDR and significant values are taken as <0.05. As the samples are small and the random sets can alter the final results, we performed the analysis 100 times and at the end count the number of times the difference was found significant.

Enrichment analysis of the interaction between PTMs and protein domains

For every pair of PTM types with certain degree of co-evolution (red lines at Figure 3B), we extracted the Interpro domains (Hunter *et al*, 2009) with one of the modifications inside. We used Interpro domains instead of more curated databases, such as SMART (Letunic *et al*, 2012), to be able to include a wider definition of globular domains. We also measured the number of times that a particular modification was found inside the same domains when it is present in isolation or co-evolved with any other PTM type. We performed an enrichment analysis using Fisher exact test for each domain to test the difference between these two ratios. *P*-values were adjusted by FDR for the whole set of analysis and values <0.05 were taken as significant.

Extracting short-linear motifs enriched in sets of proteins with co-evolving PTM pairs

We performed an enrichment analysis of short-linear motifs, of 3–10 residues, for every set of proteins with a particular pair of co-evolving PTMs using the SliMFinder (Davey *et al*, 2010) software. Default parameters were used for this analysis. The extracted linear motifs were compared using CompariMotif (Davey *et al*, 2007) software to known motifs belonging to all databases available in SlimFinder.

Identification of functions enriched in proteins with co-evolving methylated and phosphorylated residues and DLF motif

We performed a functional enrichment analysis with the proteins having crosstalking methylated and phosphorylated residues and the DLF short-linear motif using the FatiGO software (Al-Shahrour *et al*, 2004) available within the Babelomics suit of tools (Medina *et al*, 2010). We used Gene Ontology terms as protein annotation and chose an adjusted (by FDR) *P*-value of 0.05 as the threshold for significance.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

PM was partially funded by an EOI fellowship from Spanish Ministry of Science and Innovation and a Marie Curie IEF fellowship (VII Framework Program). This work was supported by EMBL.

Author contributions: PB conceived the project. PM, LP, VN and PB designed the analyses. PM, LP, FD and VN performed the analyses. AG, MH and RK participated in advice and discussion. PB, VN and PM wrote the manuscript. All authors commented and agreed on the manuscript.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Al-Shahrour F, Diaz-Uriarte R, Dopazo J (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20:** 578–580

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene Ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet* **25:** 25–29

Benayoun BA, Veitia RA (2009) A post-translational modification code for transcription factors: sorting through a sea of signals. *Trends Cell Biol* **19:** 189–197

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2010) GenBank. *Nucleic Acids Res* **38:** D46–D51

Bignone PA, King MDA, Pinder JC, Baines AJ (2007) Phosphorylation of a threonine unique to the short C-terminal isoform of betaII-spectrin links regulation of alpha-beta spectrin interaction to neuritogenesis. *J Biol Chem* **282:** 888–896

Biswas AK, Noman N, Sikder AR (2010) Machine learning approach to predict proteinphosphorylation sites by incorporating evolutionary information. *BMC Bioinformatics* **11:** 273

Boekhorst J, van Breukelen B, Heck A, Snel B (2008) Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes. *Genome Biol* **9:** R144

Brooks CL, Gu W (2003) Ubiquitination, phosphorylation and acetylation: the molecular basis for p53 regulation. *Curr Opin Cell Biol* **15:** 164–171

Castellino FJ, Ploplis VA, Zhang L (2008) gamma-Glutamate and beta-hydroxyaspartate in proteins. *Methods Mol Biol (Clifton, NJ)* **446:** 85–94

Chen A, Wang P-Y, Yang Y-C, Huang Y-H, Yeh J-J, Chou Y-H, Cheng J-T, Hong Y-R, Li SSL (2006) SUMO regulates the cytoplasmonuclear transport of its target protein Daxx. *J Cell Biochem* **98:** 895–911

Chen L-F, Williams SA, Mu Y, Nakano H, Duerr JM, Buckbinder L, Greene WC (2005) NF-kappaB RelA phosphorylation regulates RelA acetylation. *Mol Cell Biol* **25:** 7966–7975

Chen SC-C, Chen F-C, Li W-H (2010) Phosphorylated and nonphosphorylated serine and threonine residues evolve at different rates in mammals. *Mol Biol Evol* **27:** 2548–2554

Chica C, Labarga A, Gould CM, López R, Gibson TJ (2008) A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. *BMC Bioinformatics* **9:** 229

Choudhary C, Kumar C, Gnad F, Nielsen ML, Rehman M, Walther TC, Olsen JV, Mann M (2009) Lysine acetylation targets protein

complexes and co-regulates major cellular functions. *Science (New York, NY)* **325:** 834–840

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science (New York, NY)* **311:** 1283–1287

Cohen P (2000) The regulation of protein function by multisite phosphorylation—a 25 year update. *Trends Biochem Sci* **25:** 596–601

Cover TM, Thomas JA (1991) *Elements of Information Theory*. New York: John Wiley & Sons

Dalrymple BP, Kongsuwan K, Wijffels G, Dixon NE, Jennings PA (2001) A universal protein-protein interaction motif in the eubacterial DNA replication and repair systems. *Proc Natl Acad Sci USA* **98:** 11627–11632

Danielsen JMR, Sylvestersen KB, Bekker-Jensen S, Szklarczyk D, Poulsen JW, Horn H, Jensen LJ, Mailand N, Nielsen ML (2011) Mass spectrometric analysis of lysine ubiquitylation reveals promiscuity at site level. *Mol Cell Proteomics: MCP* **10:** 003590

Darnell GA, Antalis TM, Johnstone RW, Stringer BW, Ogbourne SM, Harrich D, Suhrbier A (2003) Inhibition of retinoblastoma protein degradation by interaction with the serpin plasminogen activator inhibitor 2 via a novel consensus motif. *Mol Cell Biol* **23:** 6520–6532

Davey NE, Edwards RJ, Shields DC (2007) The SLiMDisc server: short, linear motif discovery in proteins. *Nucleic Acids Res* **35:** W455–W459

Davey NE, Haslam NJ, Shields DC, Edwards RJ (2010) SLiMFinder: a web server to find novel, significantly over-represented, short protein motifs. *Nucleic Acids Res* **38:** W534–W539

Deribe YL, Pawson T, Dikic I (2010) Post-translational modifications in signal integration. *Nat Struct Mol Biol* **17:** 666–672

Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, Diella F (2010) Phospho.ELM: a database of phosphorylation sites—update 2011. *Nucleic Acids Res* **39:** D261–D267

Gnad F, Gunawardena J, Mann M (2010) PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res* **39:** D253–D260

Gnad F, Ren S, Cox J, Olsen JV, Macek B, Oroshi M, Mann M (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol* **8:** R250

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biol* **59:** 307–321

Gupta R, Birch H, Rapacki K, Brunak S, Hansen JE (1999) O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res* **27:** 370–372

Harrison BC, Huynh K, Lundgaard GL, Helmke SM, Perryman MB, McKinsey TA (2010) Protein kinase C-related kinase targets nuclear localization signals in a subset of class IIa histone deacetylases. *FEBS Lett* **584:** 1103–1110

Holt LJ, Tuch BB, Villén J, Johnson AD, Gygi SP, Morgan DO (2009) Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science (New York, NY)* **325:** 1682–1686

Hornbeck PV, Chabra I, Kornhauser JM, Skrzypek E, Zhang B (2004) PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* **4:** 1551–1561

Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T (2007) The human phylome. *Genome Biol* **8:** R109

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R *et al* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res* **37:** D211–D215

Hunter T (2007) The age of crosstalk: phosphorylation, ubiquitination, and beyond. *Molecular Cell* **28:** 730–738

Huynen M (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res* **10:** 1204–1210

Jensen LJ, Jensen TS, de Lichtenberg U, Brunak S, Bork P (2006) Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature* **443:** 594–597

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C (2009) STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* **37:** D412–D416

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M *et al* (2009) Human protein reference database—2009 update. *Nucleic Acids Res* **37:** D767–D772

Kontaki H, Talianidis I (2010) Cross-talk between post-translational modifications regulates life or death decisions by E2F1. *Cell Cycle* **9:** 3836–3837

Korber B, Farber R, Wolpert D, Lapedes AS (1993) Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc Natl Acad Sci* **90:** 7176–7180

Kumar M, Balaji PV (2011) Comparative genomics analysis of completely sequenced microbial genomes reveals the ubiquity of N-linked glycosylation in prokaryotes. *Mol Biosyst* **7:** 1629–1645

Landry CR, Levy ED, Michnick SW (2009) Weak functional constraints on phosphoproteomes. *Trends Genet* **25:** 193–197

Lange A, Mills RE, Lange CJ, Stewart M, Devine SE, Corbett AH (2007) Classical nuclear localization signals: definition, function, and interaction with importin alpha. *J Biol Chem* **282:** 5101–5105

Latham JA, Dent SYR (2007) Cross-regulation of histone modifications. *Nature Struct Mol Biol* **14:** 1017–1024

Lee T-Y, Huang H-D, Hung J-H, Huang H-Y, Yang Y-S, Wang T-H (2006) dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res* **34:** D622–D627

Letunic I, Doerks T, Bork P (2012) SMART7: recent updates to the protein domain annotation resource. *Nucleic Acids Res* **40:** D302–D305

Levy D, Kuo AYAJ, Chang Y, Schaefer U, Kitson C, Cheung P, Espejo A, Zee BM, Liu CL, Tangsombatvisit S, Tennen RI, Kuo AY, Tanjing S, Cheung R, Chua KF, Utz PJ, Shi X, Prinjha RK, Lee K, Garcia BA *et al* (2011) Lysine methylation of the NF-κB subunit RelA by SETD6 couples activity of the histone methyltransferase GLP at chromatin to tonic repression of NF-κB signaling. *Nature Immunol* **12:** 29–36

Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB (2003) Protein disorder prediction: implications for structural proteomics. *Structure (London, England: 1993)* ***11*** 1453–1459

Lu Z, Cheng Z, Zhao Y, Volchenboum SL (2011) Bioinformatic analysis and post-translational modification crosstalk prediction of lysine acetylation. *PLoS ONE* **6:** e28228

Malik R, Nigg EA, Körner R (2008) Comparative conservation analysis of the human mitotic phosphoproteome. *Bioinformatics (Oxford, England)* **24:** 1426–1432

Mani KM, Lefebvre C, Wang K, Lim WK, Basso K, Dalla-Favera R, Califano A (2008) A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Mol Syst Biol* **4:** 169

Marcet-Houben M, Gabaldón T (2011) TreeKO: a duplication-aware algorithm for the comparison of phylogenetic trees. *Nucleic Acids Res* **39:** e66

Martin LC, Gloor GB, Dunn SD, Wahl LM (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics (Oxford, England)* **21:** 4116–4124

Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, Conesa A, Tárraga J, Pascual-Montano A, Nogales-Cadenas R, Santoyo J, García F, Marbà M, Montaner D, Dopazo J (2010) Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res* **38:** W210–W213

Mills IG, Praefcke GJK, Vallis Y, Peter BJ, Olesen LE, Gallop JL, Butler PJG, Evans PR, McMahon HT (2003) EpsinR: an AP1/clathrin

interacting protein involved in vesicle trafficking. *J Cell Biol* **160:** 213–222

Minguez P, Götz S, Montaner D, Al-Shahrour F, Dopazo J (2009) SNOW, a web-based tool for the statistical analysis of protein-protein interaction networks. *Nucleic Acids Res* **37**(Web Server issue)**:** W109–W114

Muller J, Creevey CJ, Thompson JD, Arendt D, Bork P (2010a) AQUA: automated quality improvement for multiple sequence alignments. *Bioinformatics (Oxford, England)* **26:** 263–265

Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, Kuhn M, Powell S, von Mering C, Doerks T, Jensen LJ, Bork P (2010b) eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res* **38:** D190–D195

Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, Gibson TJ, Lewis J, Serrano L, Russell RB (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol* **3:** e405

Oppermann FS, Gnad F, Olsen JV, Hornberger R, Greff Z, Kéri G, Mann M, Daub H (2009) Large-scale proteomics analysis of the human kinome. *Mol Cell Proteom* **8:** 1751–1764

Perrotta S, del Giudice EM, Iolascon A, De Vivo M, Pinto DD, Cutillo S, Nobili B (2001) Reversible erythrocyte skeleton destabilization is modulated by beta-spectrin phosphorylation in childhood leukemia. *Leukemia* **15:** 440–444

Rodriguez MS, Dargemont C, Hay RT (2001) SUMO-1 conjugation in vivo requires both a consensus modification motif and nuclear targeting. *J Biol Chem* **276:** 12654–12659

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T *et al* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **39:** D38–D51

Seet BT, Dikic I, Zhou M-M, Pawson T (2006) Reading protein modifications with interaction domains. *Nature Rev Mol Cell Biol* **7:** 473–483

Shimazu T, Horinouchi S, Yoshida M (2007) Multiple histone deacetylases and the CREB-binding protein regulate pre-mRNA 3′-end processing. *J Biol Chem* **282:** 4470–4478

Shukla A, Chaurasia P, Bhaumik SR (2009) Histone methylation and ubiquitination with their cross-talk and roles in gene expression and stability. *Cell Mol Life Sci* **66:** 1419–1433

Spilianakis C, Papamatheakis J, Kretsovali A (2000) Acetylation by PCAF enhances CIITA nuclear accumulation and transactivation of major histocompatibility complex class II genes. *Mol Cell Biol* **20:** 8489–8498

Sudol M (1998) From Src Homology domains to other signaling modules: proposal of the 'protein recognition code'. *Oncogene* **17:** 1469–1474

Tan CSH, Bader GD (2012) Phosphorylation sites of higher stoichiometry are more conserved. *Nature Methods* **9:** 317–317

Tan CSH, Bodenmiller B, Pasculescu A, Jovanovic M, Hengartner MO, Jørgensen C, Bader GD, Aebersold R, Pawson T, Linding R (2009) Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Sci Signal* **2:** ra39

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science (New York, NY)* **278:** 631–637

The UniProt Consortium (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* **38:** D142–D148

van Noort V, Seebacher J, Bader S, Mohammed S, Vonkova I, Betts MJ, Kühner S, Kumar R, Maier T, O'Flaherty M, Rybin V, Schmeisky A, Yus E, Stülke J, Serrano L, Russell RB, Heck AJ, Bork P, Gavin AC (2012) Cross-talk between phosphorylation and lysine acetylation in a genome-reduced bacterium. *Mol Syst Biol* **8:** 571

Vethantham V, Rao N, Manley JL (2008) Sumoylation regulates multiple aspects of mammalian poly(A) polymerase function. *Genes Dev* **22:** 499–511

Wang Z, Udeshi ND, Slawson C, Compton PD, Sakabe K, Cheung WD, Shabanowitz J, Hunt DF, Hart GW (2010) Extensive crosstalk between O-GlcNAcylation and phosphorylation regulates cytokinesis. *Sci Signal* **3:** ra2

Weinert BT, Wagner SA, Horn H, Henriksen P, Liu WR, Olsen JV, Jensen LJ, Choudhary C (2011) Proteome-wide mapping of the drosophila acetylome demonstrates a high degree of conservation of lysine acetylation. *Sci Signal* **4:** ra48–ra48

Yang X-D, Tajkhorshid E, Chen L-F (2010) Functional interplay between acetylation and methylation of the RelA subunit of NF-kappaB. *Mol Cell Biol* **30:** 2170–2180

Yang X-J (2005) Multisite protein modification and intramolecular signaling. *Oncogene* **24:** 1653–1662

Zielinska DF, Gnad F, Wiśniewski JR, Mann M (2010) Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell* **141:** 897–907

